Search Upgrade Project Report
November 19, 2003

Members of the Computing Division, the Fermilab Office of Public Affairs and the D0 and CDF Experiments discussed various aspects of search software and evaluated the search software currently in use by Fermilab as part of a Search Working Group (SWG) in September, 2003. In addition to upgrading the software (formerly Inktomi, now Verity Ultraseek) to the latest release, members of the Working Group were asked to provide requirements of the software itself as well as requirements of what content needed to be searchable. What follows is a report of the Search Working Group's findings.

A. Project Definition

http://wwwserver2.fnal.gov/cfdocs/projectsdb/projdetail.cfm?ProjectID=138

B. Members

Marcia Teckenbrock, Project Leader, CD/CDO Projects and Outreach
Andy Beretvas, CD/CDF Computing & Analysis
Kevin Hill, CD/CSS/CSI
Qizhong Li, CD/D0 Computing & Analysis
Kevin Munday, Xeno Media (for the Fermilab Office of Public Affairs)

Additional Input from:

Ruth Pordes
Glenn Cooper, CDF Experiment
Alan Jonckheere, D0 Experiment

C. Software Requirements and Issues

The main points the SWG needed to address were:

1. The Fermilab Office of Public Affairs has public content that needs to be indexed on a regular basis.

   In particular, the *Fermilab Today* newsletter, which is published daily, should be indexed every day. Tests showed that *Fermilab Today* was indexed by the Google search engine only on a weekly basis, which is not sufficient to meet the needs of the Office of Public Affairs.

   The content managed by the Office of Public Affairs, in collaboration with Xeno Media, is important and relevant to all those working at Fermilab, as well as those seeking information about Fermilab, including government staff. With this in mind, daily indexing is a major requirement that should not be overlooked.

2. The CDF Experiment has private content served on the Web. Spokespeople for CDF did not wish private content to be indexed, due to the availability of page titles and other small amounts of information in the search results.

3. The D0 Experiment also has private content at some level they might wished indexed at a later date. At this point in time, spokespeople for D0 do not wish any of the private content to be indexed either, due to the availability of page titles and small amounts of information in the search results.

4. Don Petravick asked about having the capability to index Listserv archives at the Search Briefing. Although this is possible with the Ultraseek software, discussion with CSI management advised against it. Firstly, the listserv archives have their own search function., although searching across more than one archive at a time is not allowed. Secondly, many of the lists are closed, and creating and maintaining such an index would be tedious. Furthermore, CSI has not had any other requests to index Listserv archives.

D. Content Issues

1. CDF is planning to investigate search engines for use on their own. The CDF main web server should continue to be indexed by the Fermilab search engine, as it has been.

2. The CMS Experiment is planning on using the free Google search service for their content.

3... D0 provided a list of three new servers which should be included in their public search. DZero also would like to investigate indexing relevant content on other servers off-site.

4. Work by the SWG did not determine further content to index within the Computing Division.

5. The Beams Division has been contacted regarding recommendations they may have. They will respond, but have not to date.

E. Indexing Issues

1. Testing showed that search results could be improved by a low-level of effort from management or others familiar with the content of their web sites.

2. Evaluating search results from time-to-time would make search results better, as impertinent content could be pulled from the indices. Commitment by management to ensure that the appropriate people would provide feedback to the search administrator would be ideal.

3. A common interface for searching across both Google and the on-site search engine is desired by CDO-Projects. Xeno Media has expressed interest in developing such an interface, using scripting technology. Partial use of existing Purchase Orders should make this possible.

F. Recommendations

1. The Search Working Group recommends that the Lab continue to use the current Verity Ultraseek search software in conjunction with Google to provide optimal search results. Clearly however, the cost of the Ultraseek maintenance should be carefully considered in conjunction with this recommendation.

2. It is important to have feedback from the Experiments, Divisions and Departments from time-to-time to determine if the appropriate content is being indexed. This is something that would most likely need to come as a directive from Upper Management. In addition, it is possible to have different collections administered by different people. It may be useful to make use of this feature.

3. There is a free, Java-based product called Jakarta Lucene that is being planned for use in an application being developed by CEPA members. This software would require setup and administration by a Java programmer. It may be worthwhile to investigate other free search engines over the next several months (Our maintenance contract with Verity ends 30 August, 2004).

G. Justification

1. Google's free service does not meet an important requirement by the Fermilab Office of Public Affairs: indexing content in a timely manner.

2. Google's free service does not allow as much control over the information that is indexed as an on-site search engine. Password restrictions and robots.txt files can prevent indexing of unwanted content, however, one cannot control how often or exactly which content is indexed when using an external search engine.

3. Testing showed that Google handled relevance ranking better than Ultraseek. Results using the on-site search engine could be improved by feedback from users and content providers.

H. Impact

1. Support for Verity Ultraseek
    a. The latest version of Verity Ultraseek must be moved into production. We have purchased an extra machine for development, but it was not needed.
    b. Support for Verity
        i. Maintenance: ~$25,000/year
        ii. Employee Effort: approximately two workdays a month for 1 FTE at most
        iii. Management of collections may be offloaded to those serving the content.

I. Appendices

1. Search engine evaluation form

2. Search Upgrade Project Plan

3. Project Definition

4. Current Sites Using On-site Search Engine

Search Engine Evaluation Form

Which index are you searching? Choose only one index per form:

___ Computing Division
___ Fermilab (www.fnal.gov only)
All Fermilab
___ Fermilab Today
___ DZero
___ HEPIC
___ FermiNews
___ CDF

Search Term	Google Ranking	# Results Google	Relevance Rating : Google Inktomi (1=low, 5=high)	Ultraseek Ranking	#Results Ultraseek (new software)	Relevance Rating : Ultraseek (1=low, 5=high)	#Results Inktomi	Relevance Rating	Comments

ID	Activity Desc.	Duration	Activity Type	Actual Start	Actual Finish	Early Dates	Progress Type	Progress Value
1	Identify Members of Working Group	10d	ASAP	15-Sep-03	25-Sep-03	15Sep03 25Sep03	Complete	100%
2	Upgrade Search Software	3d	ASAP	23-Sep-03	29-Sep-03	23Sep03 29Sep03	Percent Complete	100%
3	Guidance statement document to understand indexing	3d	ASAP			03Dec03 05Dec03	Percent Complete	10%
4	Complete list of requirements/test points required of sw	3d	ASAP			05Nov03 06Nov03	Percent Complete	40%
5	Identifying what content needs to be indexed	38d	ASAP	1-Oct-03		01Oct03 18Nov03	Percent Complete	50%
6	Begin testing software	6d	ASAP	28-Oct-03		29Oct03 04Nov03	Percent Complete	20%
7	Report findings/begin writing report	1d	ASAP			19Nov03 19Nov03	Planned	0
8	Submit finished report detailing findings to management	7d	ASAP			19Nov03 26Nov03	Planned	0
9	Report feedback from Div/Dept heads & experiments	10d	ASAP			18Dec03 18Dec03	Planned	0

Search Upgrade Project Definition

Responsible OU	Task Number	Stakeholders
		CD, Public Affairs, DO, CDF, CMS

Leader(s)	Participant(s)	Effort
M Teckenbrock		

Start Date	End Date	Status
		Active

Deliverables

A report will be generated by the Search Working Group, detailing its findings. The report will be used to determine whether we continue using the Verity software or look for a new search solution

Description

We plan to upgrade our search software from Inktomi Search 4.5.0 to Verity Ultraseek 5.1.0. We will test the software to determine whether it meets our needs and form a working group to evaluate.

Plan

Identify Members of Working Group (Sept 25); Upgrade Search Software (Dec 29); Guidance statement document to understand indexing (Dec 5); Complete list of requirements/test points required of sw (Nov 19); Identify what content needs to be indexed (Oct 1); Begin testing software (Oct 28); Report findings/begin writing report (Nov 19); Submit finished report detailing findings to management (Nov 26); Report feedback from Div/Dept heads & experiments (Dec 18)

Schedule

Issues

Comments

Project URL

Current Sites Using On-Site (Verity Ultraseek) Search Engine

Site Indexed	Contact	Search Page Location	Comments	Uses on-site Search
CDF Experiment	Glenn Cooper gcooper@fnal.gov	http://www-cdf.fnal.gov/cdfsearch.html	Will investigate own search engine for private content.	√
D0 Experiment	Alan Jonckheere d0web-support@fnal.gov, Stu Fuess fuess@fnal.gov	http://www.hep.net/search/d0.fnal.gov (maintained by Marcia)		√
DOE- Office of High Energy Physics	Marsha Marsden Marsha.Marsden@SCIENCE.DOE.GOV	http://doe-hep.hep.net/search.html (maintained by Marcia)		√
FNAL – CD	Marcia Teckenbrock marcia@fnal.gov	http://cddocs.fnal.gov/cfdocs/productsDB/docs.html (maintained by CDO-Projects)		√
FNAL – Office of Public Affairs, (including Fermilab Today and FermiNews)	Kevin Munday munday@xenomedia.com	http://www.fnal.gov/pub/search/index.html		√
HEPIC	Marcia Teckenbrock marcia@fnal.gov	http://www.hep.net/search/hepic.html (maintained by Marcia)		√
SELEX Experiment	Peter Cooper pcooper@fnal.gov	http://www.hep.net/search/selex.html (maintained by Marcia)		√